Comput. & Ops. Res. Vol. 12, No. 6, pp. 509–516, 1985
Printed in Great Britain

0305-0548/86 $3.00 + .00

LOCATING AN n-SERVER FACILITY IN A STOCHASTIC ENVIRONMENT

SAMUEL S. CHIU†
Department of Engineering-Economic Systems, Stanford University, Stanford, CA 94305, U.S.A.

and

RICHARD C. LARSON‡
Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A.

Scope and Purpose—Planners for spatially distributed service systems must confront the problem of locating their service facilities. In this paper we address the situation in which one such facility is to be located to service an entire region. The facility is assumed to house a given number of mobile service units (e.g. emergency repair vehicles, ambulances), any one of which can be dispatched to handle an on-scene service request. Both the arrival times and service requirements (i.e. service times) of service requests are probabilistic, representing an advance over earlier models. Queueing delays are not allowed, so it is assumed that any service requests that occur while all service units are busy are handled (at a cost) by some back-up service system. While our model and analysis differ from traditional deterministic location models, we show that the optimum location for the facility is the same as that obtained from a well-known deterministic model.

Abstract—We consider a demand-responsive service system in which n mobile units (servers) are garaged at one facility. Service demands arrive in time as a homogenous Poisson process, but are located over the service region according to an arbitrary probability law. Given a random service demand, either (1) a mobile unit is dispatched to the demand's location to provide on-scene service or (2) the demand is lost (i.e. it is handled by some back-up system). The resultant queueing system is an $M/G/n$ loss system operating in steady state. The objective is to locate the garage facility so that the average cost of response is minimized, where the cost of response is a weighted sum of mean travel time to a random serviced demand and the cost of a lost demand, the weights being the respective probabilities of occurrence. We show that the optimum facility location reduces to Hakimi's well-known minisum location.

1. INTRODUCTION

The motivation of this paper comes from consideration of facility location problems in which spatially dispersed "customers" receive service from mobile "servers," garaged at one or more "home locations" while not servicing customers. In the public domain, examples include ambulance services and fire departments. In the private domain, important examples are emergency repair services and certain home delivery services.

In many classical locational decisions, minimization of average system-wide travel time has been the sole objective. This decision criterion implies that servers are available at all times and each call for service by a customer is answered immediately by the nearest (always available) server. Under such a congestion free environment, average travel time is a good measure of system performance. In an urban environment, service-oriented systems are plagued with uncertainties: temporal and spatial uncertainties as to the arrivel time and location, respectively, of the next customer; service time uncertainties as to the amount of time required to complete a service; travel time uncertainties as to the fluctuations of transit

†Professor Samuel S. Chiu received his Ph.D. from MIT in Operations Research and has been with the Engineering-Economic Systems Department at Stanford University since 1982. He is a recipient of the 1984 Presidential Young Investigator Award administered by the National Science Foundation. His research interests include quantitative methods in urban service systems, locational theory, and incentive theory with application to power pooling. He coauthored a chapter in *Discrete Location Theory* coedited by Francis and Mirchandani. He is a consultant with Arthur D. Little, Inc., Cambridge, Massachusetts.

‡Richard C. Larson is Professor of Electrical Engineering and Urban Studies and Planning at MIT, Cambridge, MA 02139. He is also Codirector of the MIT Operations Research Center. He received the S.B., S.M., and Ph.D. degrees from MIT in 1965, 1967, and 1969, respectively. His most recent book is *Urban Operations Research* (Amedeo R. Odoni, coauthor), Prentice–Hall (1981). His papers have appeared in *Management Science, Operations Research, Transportation Science, Networks,* and other journals. He is a member of TIMS, ORSA, AAAS, and IEEE. His current research interests include the psychology of queueing and intelligent computer-aided dispatch systems.

time. As a consequence of such uncertainties, in urban emergency service settings where system utilization (i.e. the fraction of time that a server is busy) is high, customers often find all servers unavailable at the time a service request is initiated. Such a customer could wait in a queue until an appropriate server is free, or the request for service could be lost at a cost (e.g. being served by a distant or more expensive back-up unit). It is the latter case, the case of "loss," that we will analyze in this paper.

Most previous queueing considerations in this context relied on the assumption of exponentially distributed service times. Such an assumption allows one to model the system as a continuous time Markov process. Mirchandani, Silva, and Visocki [10] study the resultant Markov model for a network with two nodes. Berman and Larson [1] extend the Markov analysis to an arbitrary network having n servers, any combination of which may be unavailable (i.e. busy) at a given moment. They obtain Hakimi-type [5] nodal optimality results; i.e. there always exists a set of optimal locations on the nodes of a network. They assume that the state probabilities of server status (free or busy) remain fixed for local changes of service facility location (except when the change of facility locations results in changes of server assignment policies). This approximation is good only when on-scene service time is much larger than travel time.

In many applications of interest, the total service time (defined to be the sum of travel time and on-scene [and perhaps related off-scene] time), is dominated by travel time. The resultant distribution for total service time is (i) non-Markovian [i.e. not exponential] and (ii) dependent on the location(s) of the server(s). Allowing for such complications, Berman, Larson, and Chiu [2] prove the nodal optimality result for the one server loss case (demand is lost at a cost when the lone server is busy). They do not make any restrictive assumption about service time distribution nor do they assume constancy of the state probabilities upon small changes of facility locations.

In this paper, we extend the model of Berman, Larson, and Chiu [2] in two ways: (i) we wish to locate a single facility for n servers; and (ii) we allow any topological structure, distance metric, and demand distribution in the loss model. When our result is applied to a network structure with only nodal demands, it is both more general and more restrictive than the problem studied in Berman and Larson [1]. It is more general because we make no assumptions on service time distribution and we allow the steady state probabilities to vary continuously as we change facility location. It is more restrictive because we are locating one single home for all servers.

Our objective is to locate a single garage facility for n servers, which will minimize the expected weighted cost of travel time and the cost of lost customers. We will call such a location an n-server-single-facility-loss median (n-SFLM). We will show that the standard average travel time minimizing location coincides with our n-SFLM under any topological setting with any demand distribution over the region of interest. The situations include continuous link demands on a general network (see Chiu [4]), location problems on an Euclidean plane, on a plane with L_1 (or L_p) metric (see Larson and Sadiq [8], and Odoni and Sadiq [11]). The cost of "loss" is, curiously enough, only constrained to be nonnegative. Section 3 contains a discussion of the behavior of our objective function when the underlying topology is a network with discrete nodal demands. We end this paper with a discussion of the difficulties one encounters in generalizing the loss model we study.

For general background on facility location problems, we refer the reader to Francis and White [6], Thisse and Zoller [13], and Francis and Mirchandani [9].

2. MODEL FORMULATION AND RESULTS

Consider a region R in which customers' requests for service are generated as a time homogeneous Poisson process with rate λ. The location of such a service request is governed by some probability distribution function over the region of interest. A single server, if at least one of the total pool of n is available, will travel from the home location of the n servers to the point of service request, provide the required service, return to the home location, and prepare itself for another request. Figure 1 shows the temporal sequence

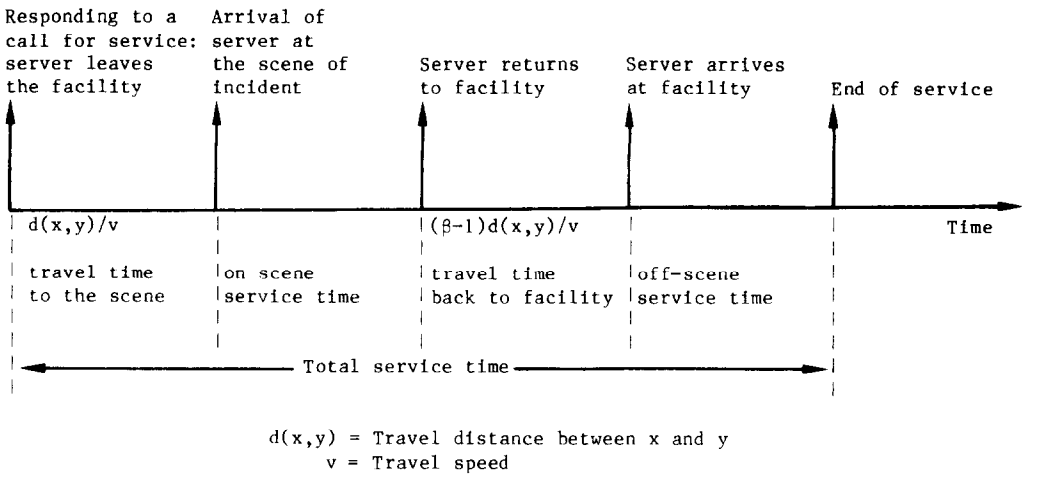

Fig. 1. Temporal sequence of a service duration for a facility at x and incident at y.

comprising a single total service time, during which the server is considered busy. A request for service is lost at a nonnegative cost Q if all n servers are busy upon its arrival. Our objective is to locate the garage facility for the n servers so that the weighted sum of mean travel time (when at least one server is idle) and the cost of loss Q (when all servers are busy) is minimized.

Before we analyze and search for the n-SFLM, we make a few remarks: (1) For a garage facility located at x, the service time distribution depends only on call location and not on system status, nor does it depend on server identity. (2) We are locating one facility for n servers. (3) The location of the facility determines not only the service time distribution for each customer location, it also determines the system's steady state probabilities. (4) For a facility located at x, the objective function takes on the following form:

$$Z(x) = [1 - P_s(x)]\bar{t}(x) + P_s(x)Q, \tag{1}$$

where

$P_s(x) \equiv$ probability of saturation (i.e. all servers busy),

$\bar{t}(x) \equiv$ expected one-way travel time,

$Q \equiv$ cost per lost customer, $Q \geq 0$.

The argument x denotes location of the facility in the region of interest. The expected travel time can be viewed as expected response time, since a customer is either served immediately or lost. (5) All servers are indistinguishable in terms of their service time distribution and hence the expected service time. To evaluate $Z(x)$, we need to know the expression for $P_s(x)$. This is readily available as the Erlang Loss Formula (Takacs [12]):

$$P_s(x) = \frac{\rho(x)^n/n!}{\sum_{i=0}^{n} \rho(x)^i/i!}, \tag{2}$$

$$\rho(x) = \lambda\bar{w} + \lambda\beta\bar{t}(x) = \lambda\bar{s}(x), \tag{3}$$

$$\bar{s}(x) = \bar{w} + \beta\bar{t}(x)$$

$= $ expected service time for one service request,

$\bar{w} = $ expected nontravel related component of total service time,

$\beta \geq 1$ accounts for round trip travel time; e.g. when $\beta = 2$, it takes as much time traveling from the facility at x to the incident location as it does on the return trip.

Thus $\bar{s}(x)$ consists of round trip travel time and other nontravel related service time.

Our optimization problem becomes

$$\underset{x \in R}{\text{Min}} \; Z(x),$$

where R is the region of interest.

We define a point x in R that minimizes $\bar{t}(x)$ as a minisum location. We will first prove a weaker version of our main result equating the minisum location with the n-SFLM.

THEOREM 1:

The n-SFLM coincides with the minisum location if $Q \geq \bar{t}(x)$ for all x in R.

Proof. Since the minisum location minimizes $\bar{t}(x)$, we need only show that

$$\frac{dZ(x)}{d\bar{t}(x)} \geq 0 \text{ for all } x \in R,$$

but

$$\frac{dZ(x)}{d\bar{t}(x)} = [1 - P_s(x)] + \lambda\beta[Q - \bar{t}(x)]\frac{dP_s(x)}{d\rho(x)} \geq 0$$

if

$$\frac{dP_s(x)}{d\rho(x)} \geq 0, \text{ since } [1 - P_s(x)], \lambda, \beta, \text{ and } [Q - \bar{t}(x)] \text{ are all nonnegative.}$$

since

$$\frac{dP_s(x)}{d\rho(x)} = \left[\sum_{i=0}^{n} \rho(x)^i/i!\right]^{-2} \frac{\rho(x)^{n-1}}{(n-1)!}\left[\sum_{i=0}^{n} \rho(x)^i\left(\frac{n-i}{i!n}\right)\right] \geq 0, \text{ our result follows.}$$

We offer the following intuitive interpretation for Theorem 1: to minimize $Z(x)$, we have to minimize $P_s(x)$, the probability of saturation. This is because the cost of loss Q is larger than the travel time $\bar{t}(x)$. Thus we would like to minimize $P_s(x)$, the chance of system saturation. $P_s(x)$ is the probability that all servers are busy and $\rho(x)$ is the utilization factor which is linear in $\bar{t}(x)$ (since $\bar{w}$ is independent of server location). Increasing $\bar{t}(x)$ increases the expected service time for each call. This, in turn, should produce more work for the servers and thus increase the system saturation probability. Therefore, a location that minimizes $\bar{t}(x)$ also minimizes $Z(x)$.

It turns out that we can relax the bound on Q and obtain the same result. The proof is direct but algebraically involved.

THEOREM 2:

The n-SFLM coincides with the minisum location for $Q \geq 0$. Proof: For simplicity, we will suppress the argument x in $P_s(x)$, $\bar{t}(x)$, etc.

$$\frac{dZ}{d\bar{t}} = (1 - P_s) - \lambda\beta\bar{t}\frac{dP_s}{d\rho} + \lambda\beta Q\frac{dP_s}{d\rho} \geq (1 - P_s) - \lambda\beta\bar{t}\frac{dP_s}{d\rho} \geq (1 - P_s) - \rho\frac{dP_s}{d\rho},$$

since $\lambda\beta Q(dP_s/d\rho) \geq 0$ from Theorem 1, and $\rho = \lambda\bar{w} + \lambda\beta\bar{t} \geq \lambda\beta\bar{t}$.

Therefore, we need only to show that $(1 - P_s) - \rho(dP_s/d\rho) \geq 0$.

$$(1 - P_s) - \rho\frac{dP_s}{d\rho}$$

$$= \left[\sum_{i=0}^{n}\rho^i/i!\right]^{-2}\left\{\underbrace{\left(\sum_{i=1}^{n-1}\rho^i/i!\right)\left(\sum_{i=1}^{n}\rho^i/i!\right)}_{A} - \underbrace{\rho\left[\frac{\rho^{n-1}}{(n-1)!}\left(\sum_{i=0}^{n}\rho^i/i!\right) - (\rho^n/n!)\left(\sum_{i=0}^{n-1}\rho^i/i!\right)\right]}_{B}\right\}.$$

After some algebraic manipulation, B is seen to be

$$B = \sum_{k=0}^{n-1}\left(\frac{n-k}{n!k!}\right)\rho^{n+k} \quad .$$

Since each term in A is positive, we need only compare terms in A that involve ρ^{n+k} with $\left(\frac{n-k}{n!k!}\right)\rho^{n+k}$ for $k = 0, 1, \ldots n-1$. We compute the coefficients of ρ^{n+k} in A to be,

for $k = 0, 1, 2, \ldots n-1$:

$$C(\rho^{n+k}) = \sum_{j=k+1}^{n}\frac{1}{j!(n-j+k)!}$$

$$= \underbrace{\frac{1}{(k+1)!(n-1)!} + \frac{1}{(k+2)!(n-2)!} + \frac{1}{(k+3)!(n-3)!} + \cdots \frac{1}{n!k!}}_{(n-k)\text{ terms}}$$

We wish to bound the denominator of each term of the above expression by $n!k!$. That is, we wish to show that $n!k! \geq (n-i)!\,(k+i)!$ for $i \leq n-k$. This is true if and only if

$$n(n-1)\ldots(n-i+1)(n-i)!k! \geq (n-i)!(k+i)(k+i-1)\ldots(k+1)k!$$

or

$n(n-1)(n-2)\ldots(n-i+1) \geq (k+i)(k+i-1)\ldots(k+1)$ after cancelling similar terms on both sides of the inequality.

Comparing the above inequality term by term, i.e. $(n-l)$ vs $(k+i-l)$, $l = 0,1,2,\ldots,i-1$, we have $n - l \geq k + i - l$ since $n - k \geq i$.

Therefore,

$$n!k! \geq (n-i)!(k+i)!$$

or the coefficient of ρ^{n+k} (in A) is

$$C(\rho^{n+k}) \geq (n-k)\left(\frac{1}{n!k!}\right).$$

Therefore,

$$A - B \geq \sum_{k=0}^{n-1}\rho^{n+k}\left[\frac{n-k}{n!k!} - \frac{n-k}{n!k!}\right] = 0.$$

Thus,

$$(1 - P_s) - \lambda\beta\bar{t}\frac{dP_s}{d\rho} \geq 0$$

and

$$\frac{dZ}{d\bar{t}} \geq 0 \text{ follows.}$$

Therefore a location minimizing $\bar{t}(x)$ also minimizes $Z(x)$.

3. SPECIALIZING TO A NETWORK TOPOLOGY

In this section, we will discuss the behavior of $Z(x)$ on a link of a general network with discrete nodal demands. Traveling between any two points on the network takes on the shortest path. In order to minimize the amount of new notation, we will state our results without proof. We refer the interested readers to Chiu [3].

It is well known that $\bar{t}(x)$ is piecewise linear and concave on a link of a general network (e.g. see Hakimi [5]), where x denotes location of the facility on the link. This results in the nodal optimality condition for a point minimizing $\bar{t}(x)$. We will define the region over which $\bar{t}(x)$ is continuously differentiable on a link as a primary region, and the boundary between two adjacent primary regions as a breakpoint. Figure 2 shows the behavior of $\bar{t}(x)$ on link (i,j) of a general network.

After explicitly expressing $\bar{t}(x)$ in terms of shortest path travel distances between demand nodes and facility at x, one can easily prove the following results (see Chiu [3] for details):

LEMMA 1

$Z(x)$ is monotone over any primary region on a link of a general network.

Since $Z(x)$ is nondifferentiable across breakpoints [as in the case of $\bar{t}(x)$], one has to evaluate the directional derivatives of $Z(x)$ at a breakpoint from both directions in order to examine the behavior of $Z(x)$ on a link. To characterize the behavior of $Z(x)$ across a breakpoint one has the following result:

LEMMA 2

The directional derivative of $Z(x)$ at a breakpoint is nonincreasing. Lemmas 1 and 2 are sufficient to guarantee that at least one minimizer of $Z(x)$ is at a nodal location.

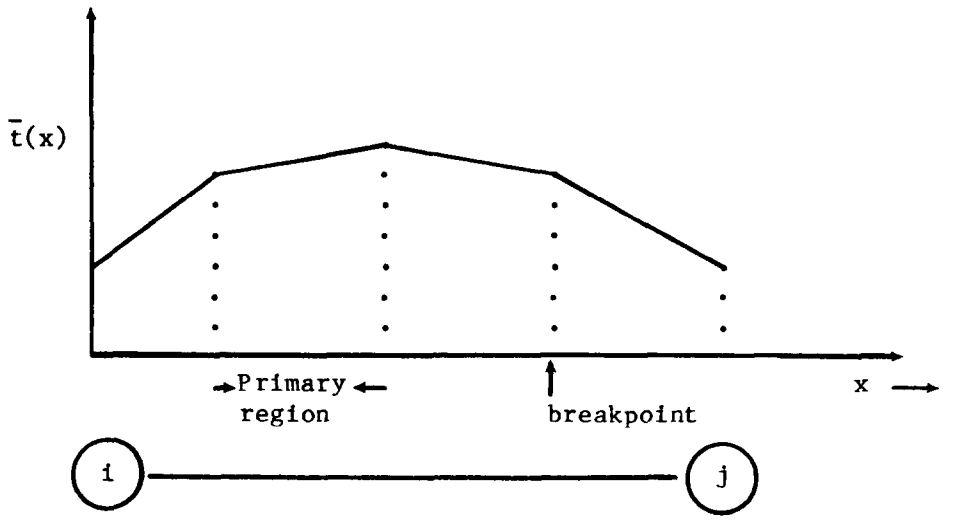

Fig. 2. $\bar{t}(x)$, primary region and breakpoint on link (i, j).

Furthermore, if we assume that $Q \geq \bar{t}(x)$ for all x on the network, one can prove the following:

LEMMA 3

If $Q \geq \bar{t}(x)$ for all x on the network, then, (i) $Z(x)$ is monotone and concave on any primary region and (ii) $Z(x)$ is concave on any link of the network.

However, Lemma 3 is not required to prove a Hakimi-type nodal optimality result. These three lemmas are quite immediate given Theorems 1 and 2, and the behavior of $\bar{t}(x)$ on a link. We will close this paper with a general discussion of the difficulties one encounters in generalizing the loss model we just analyzed.

4. CONCLUSION AND DISCUSSION

We have formulated the n-server-single-facility-loss-median problem in which we want to locate a single facility to house n servers in order to minimize the weighted average of mean travel time and the cost of lost customers. Servers are indistinguishable as far as their service time distributions are concerned. This allows us to use the Erlang Loss Formula. We do not have to evaluate the steady state probability of each combination of server status (busy or idle). Only the knowledge of the saturation probability is required. Otherwise, we have to resort to numerical methods to compute each state probability as discussed in Wolff and Wrighton [14].

We are able to prove the equivalence of the minisum location and the n-SFLM without imposing any lower bound on the value of Q (except for nonnegativity). This is very peculiar since the cost for a lost customer can take on the value zero. Intuitively, to minimize $Z(x)$, one has to strike a balance between $\bar{t}(x)$ and $P_s(x)$. When we place the facility at a very "bad" location, the expected service time increases due to the increase of travel time $\bar{t}(x)$. This will increase the saturation probability $P_s(x)$. Even if Q is zero, in which case there is no penalty for losing a customer, the increase in $P_s(x)$ [decreasing the chance that some server(s) is available] is countered by the deterioration of the travel time [and hence service time $\bar{s}(x)$]. The final mathematical analysis shows that the best location is still at a place where $\bar{t}(x)$ is minimized.

Within the framework of our model and the subsequent analysis, one can assign a server preference list (i.e. dispatch policy) at each demand point without violating any of the conclusions. Such a list allows preferences for bilingual personnel or servers familiar with local neighborhoods. Such generality also ties in with previous work of Larson [7].

We have contemplated several extensions to this loss model. The first obvious extension is that servers are allowed to reside at different locations. This modification gives rise to different service time distribution for different servers. Hence, the Erlang Loss Formula will no longer be valid. A second extension is to allow different types of calls to be handled differently by different servers and thus model $\bar{w}$, the nontravel component of total service time, to be server specific. This, of course, leads to the same difficulty. Another modification is to allow a server to respond to another service request as soon as it completes on-scene service. This modification leads to statistical dependence between successive service times, and the difficulty of deciding a dispatch policy. Available queueing theory results cannot handle the above difficulties.

Acknowledgements—This research was supported, in part, by the National Science Foundation, Grant No. 8204318-ECS, and in part by the National Science Foundation, Grant No. ECS-8307798.

REFERENCES

1. O. Berman and R. C. Larson, The median problem with congestion. *Comput. Ops. Res.* **9**, 119–126 (1982).
2. O. Berman, R. C. Larson and S. S. Chiu, Optimal server location on a network operating as an $M/G/1$ queue. *Ops. Res.* **33**, No. 4, July–August (1985).
3. S. S. Chiu, Location problems in the presence of queueing. Ph.D. dissertation, Operations Research Center, M.I.T., Cambridge, MA (1981).
4. S. S. Chiu, The minisum location problem on an undirected network with continuous link demands. Presented in the ORSA/TIMS Meeting at San Diego, 1982.

5. S. L. Hakimi, Optimal location of switching centers and absolute centers and medians of a graph. *Ops. Res.* **12**, 450–459 (1964).
6. R. L. Francis and J. A. White, *Facility Layout and Location: An Analytical Approach.* Prentice–Hall, Englewood Cliffs, N.J. (1974).
7. R. C. Larson, A hypercube queueing model for facility location and redistricting in urban emergency services. *Comput. Ops. Res.* **1**, 67–95 (1974).
8. R. C. Larson and G. Sadiq, Facility locations with the Manhattan metric in the presence of barriers to travel. *Ops. Res.* **31**, 652–669 (1983).
9. P. Mirchandani and R. Francis, *Discrete Location Theory.* Wiley, New York (1985).
10. P. B. Mirchandani, A. M. Silva and N. G. Visocki, Optimal location on a simple network with queues, Working Paper. Electrical and System Engineering Department, Rensselaer Polytechnic Institute, Troy, N.Y. (1976).
11. A. R. Odoni and G. Sadiq, Two planar facility location problems with high high speed corridors and continuous demand. *Reg. Sci. Urban Econ.* **12**, 467–484 (1982).
12. L. Takacs, *Introduction to the Theory of Queues.* Oxford University Press, New York (1962).
13. J. F. Thisse and H. G. Zoller, Eds., *Locational Analysis of Public Facilities.* North–Holland, Amsterdam (1983).
14. R. W. Wolff and C. W. Wrightson, An extension of Erlang's loss formula. *J. Appl. Prob.* **13**, 628–632 (1976).